古典中国語(漢文)の形態素解析

AUTHOR(S):

安岡, 孝一; Wittern, Christian; 守岡, 知彦; 池田, 巧; 山崎, 直樹; 二階堂, 善弘; 鈴木, 慎吾; 師, 茂樹

CITATION:

安岡, 孝一 ...[et al]. 古典中国語(漢文)の形態素解析. 東洋学へのコンピュータ利用研究セミナー 2016, 27: 3-14

ISSUE DATE:

2016-03-18

URL:

http://hdl.handle.net/2433/217946

古典中国語 (漢文) の形態素解析

安岡孝一* Christian Wittern* 守岡知彦* 池田巧*
山崎直樹† 二階堂善弘‡ 鈴木慎吾§ 師茂樹¶

1 はじめに

古典中国語 (漢文) テキストをコンピュータ処理するためには、白文 (単なる漢字の列) のままではどうにもならず、テキストを自然言語解析する必要がある。古典漢文のように、単語の間にも文の間にも区切りを持たない書写言語 (図 1) の解析では、まず、単語を認識することが必須であり、そのためには形態素解析をおこなわねばならない。

この問題に対し、われわれは、2008 年 4 月より京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの研究」を組織し、さらに 2013 年 4 月より京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの応用研究」を組織して、古典漢文に対する形態素解析の研究をおこなってきた。この研究の成果については、これまでも様々な形で公表 [1-13] してきたが、それらを本稿でざっと概観し、まとめておくことにする。


○上用大將王恢議遣恢等將兵匿馬邑旁谷中陰使聶壹誘匈奴入塞而擊之單于覺而去自是絕和親攻當路塞唐蒙上書請通南夷拜蒙中郎將將兵千人入夜郎夜郎侯聽約以爲犍爲郡又拜司馬相如爲中郎將通西夷邛筰冉駹置郡縣西至沫若水南至牂柯爲徼○徵吏民有明當世之務習先聖之術者縣次續食令與計偕詣公孫弘對

史略卷三 廿三 宴

策曰人主和德於上百姓和合於下故心和則氣和氣和則形和形和則聲和聲和則天地之和應矣策奏擢爲第一待詔金馬門齊人轅固年九十餘亦以賢良徵弘以目事之固曰公孫子務正學以言無曲學以阿世○六年初等商車○匈奴寇上谷遣將軍衛青等擊卻之○元朔元年主父偃上書諫伐匈奴嚴安亦上書及徐樂亦上書云陛下何威

図 1: 白文の例 (十八史略)

*京都大学人文科学研究所
†関西大学外国語学部
‡関西大学文学部
§大阪大学言語文化研究科
¶花園大学文学部

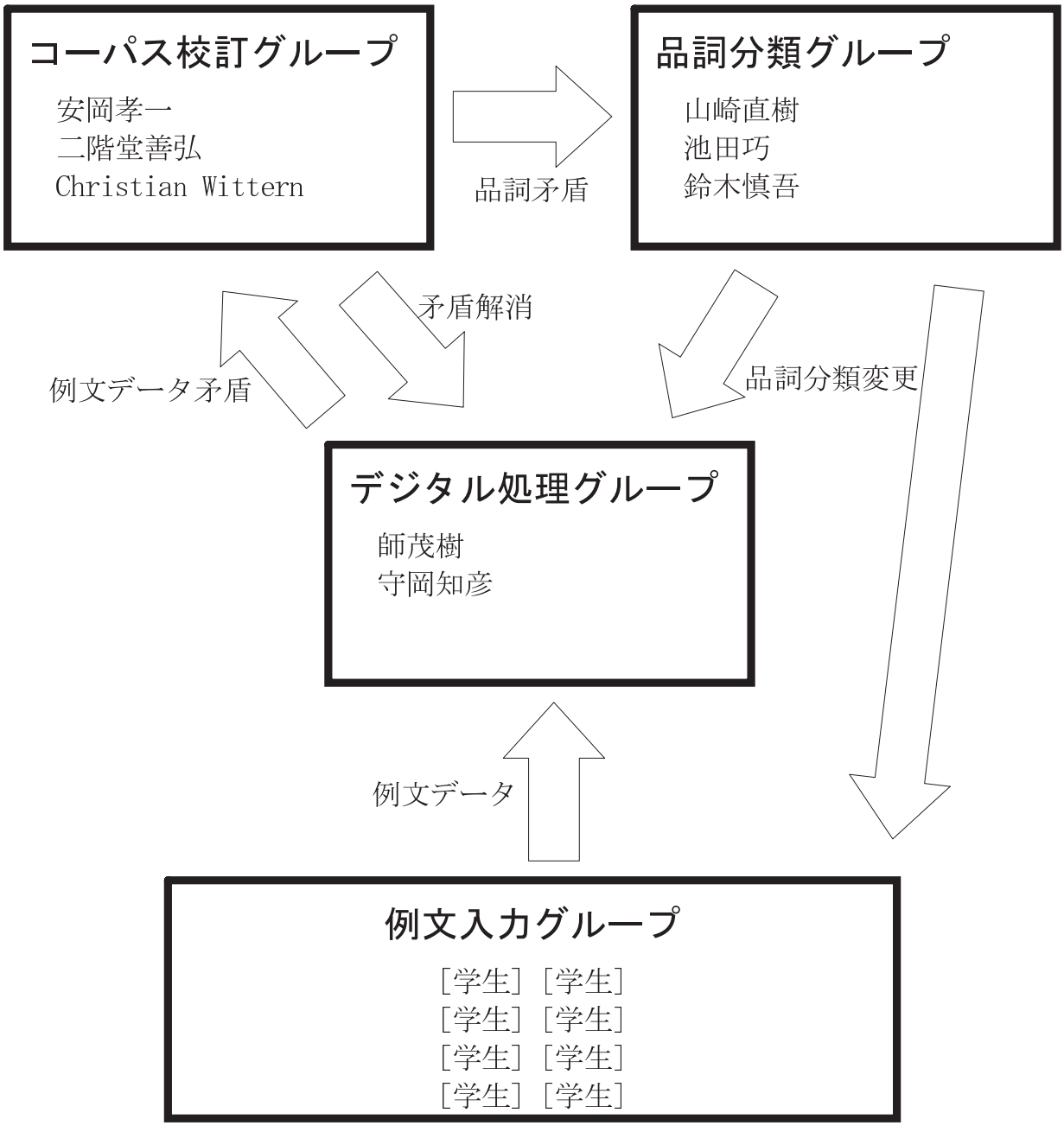


図 2: 漢文コーパス構築のための組織

2　漢文の形態素解析

漢文の形態素解析において、われわれは、MeCab というソフトウェアを用いることにした [1]。MeCab はオープンソースの形態素解析エンジンで、言語、辞書、コーパスに依存しない汎用的な設計がなされており、辞書とコーパスを準備すればいかなる言語にも対応できる、というのが売りだった。ならば、漢文 (の散文) にも MeCab を使用できるはずだ、というのが、われわれの直感だったが、われわれ以前には誰もそれを試したことがなかった。

MeCab の辞書には 4 階層の「品詞」が必要なことから、われわれは、日本語と漢文を繋ぐ「構造」の一種である訓読に着目し、返り点を「品詞」に反映させることを考えた。すなわち、訓読における返り点を、漢文の動賓構造を表しているものとみなし、動詞類に「v」という「品詞」を、賓語に「n」という「品詞」を、その他の語に「p」という「品詞」を、それぞれ、MeCab 漢文辞書の「第 1 階層の品詞」(以下「大品詞」と呼ぶ) として定めることにしたのである。次に「第 2 階層の品詞」(以下「品詞」と呼ぶ) だが、これは IPA の日本語辞書から、デッチあげてみることにした [2, 3]。「第 3 階層の品詞」(以下「意味素性」と呼ぶ) と「第 4 階層の品詞」(以下「小素性」と呼ぶ) に関しては、初期段階では付与しないことにしてみた。

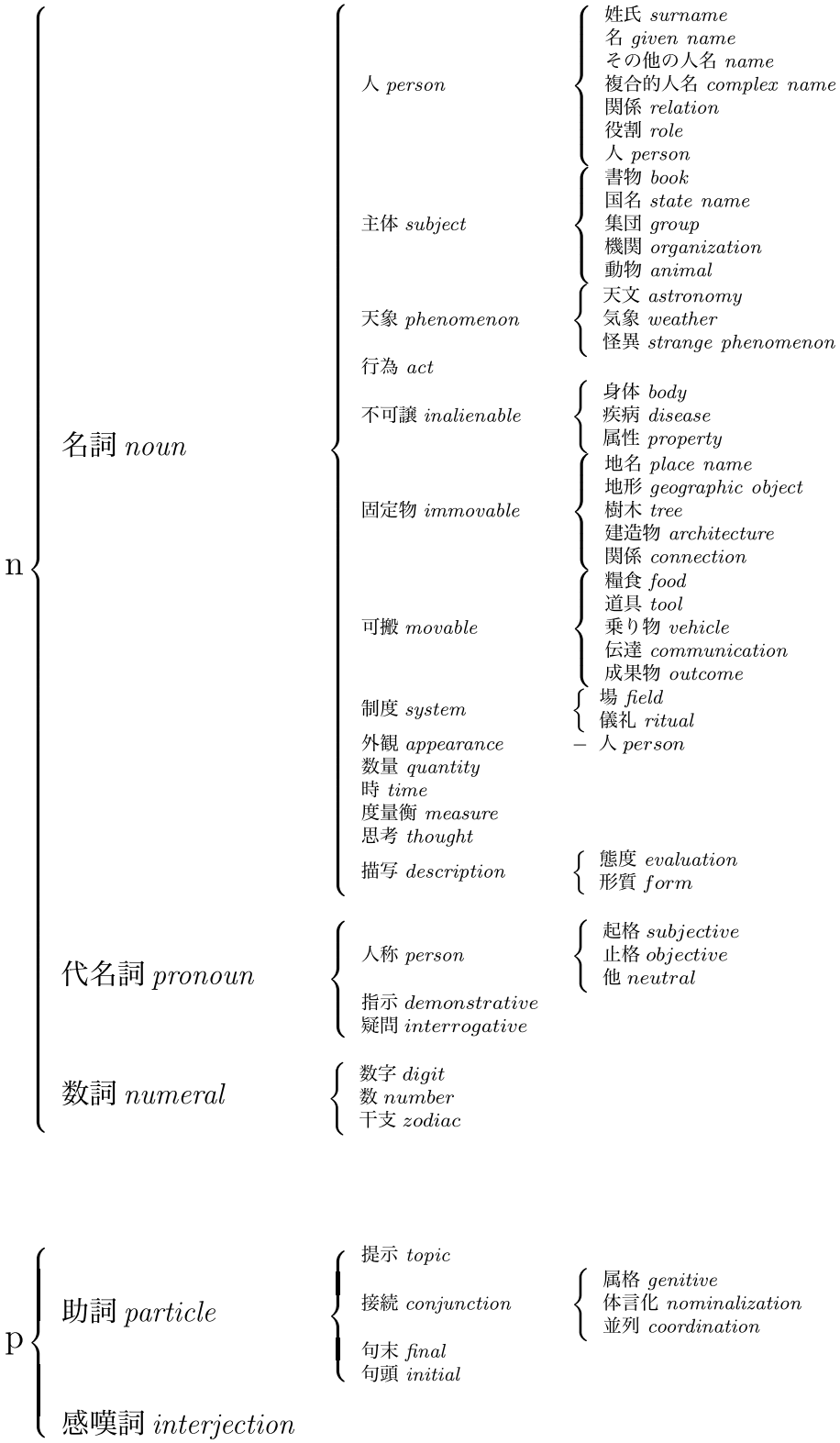
この MeCab 漢文辞書 (IPA 由来版) と、それに基づいて作った小規模な MeCab 漢文コーパスを用いて、高校教科書の漢文例や、三国志呉書列伝などの白文を、MeCab で形態素解析してみた。そうしたところ、白文を単語に区切るという点に関しては、かなり良好な結果が得られた。そこでわれわれは、例文入力グループ・デジタル処理グループ・コーパス校訂グループ・品詞分類グループの 4 グループからなる組織 (図 2) を構成し、MeCab 漢文コーパスの構築をおこなうこととした。具体的には、安岡を研究代表者とし、共同研究班の班員全員を研究分担者として、2010 年 4 月から 3 年間、科学研究費補助金基盤研究 (B) 22300087 『形態素解析のための品詞情報つき古典漢文コーパスの構築』の研究助成を受けた。

例文入力グループが MeCab 漢文コーパスを直接入力するのは、かなりの困難が予想されたことから、デジタル処理グループは、専用ツールとして、XEmacs CHISE をベースにしたコーパス入力ツールを開発した [4]。このツールは、白文を入力すると、MeCab を用いた処理をその場でおこなって、その時点での形態素解析の結果を出力する。結果に問題がなければ、そのまま漢文コーパスに反映し、もし、結果に問題があれば、入力者が手作業で訂正をおこなって、やはり漢文コーパスに反映する。たとえば、図 4 の「自立爲夜郎侯」であれば、これを正しく

```
自　　v, 副詞, 範囲, 限定,*,*, 自, 自ら, ミズカラ,*
立　　v, 動詞, 行為, 役割,*,*, 立, 立つ, タツ, 五段・タ行
爲　　v, 動詞, 行為, 役割,*,*, 爲, 為る, ナル, 五段・ラ行
夜郎　n, 名詞, 主体, 国名,*,*, 夜郎, 夜郎, ヤロウ,*
侯　　n, 名詞, 人, 役割,*,*, 侯, 侯, コウ,*
```

に訂正してから、漢文コーパスに反映する。このようなやり方で、MeCab 漢文コーパスを効率的に構築できる環境を整えた。

品詞分類グループは、コーパス校訂グループと共同で、MeCab による漢文形態素解析のための、新しい品詞体系を構築した (図 3)。この品詞体系では、大品詞を「n」「v」「p」の 3 種類とし、品詞を「名詞」「代名詞」「数詞」「動詞」「前置詞」「副詞」「助動詞」「助詞」「感嘆詞」の 9 種類として、従来の漢文文法等で見られた「形容詞」を廃止したのが特徴である。こ



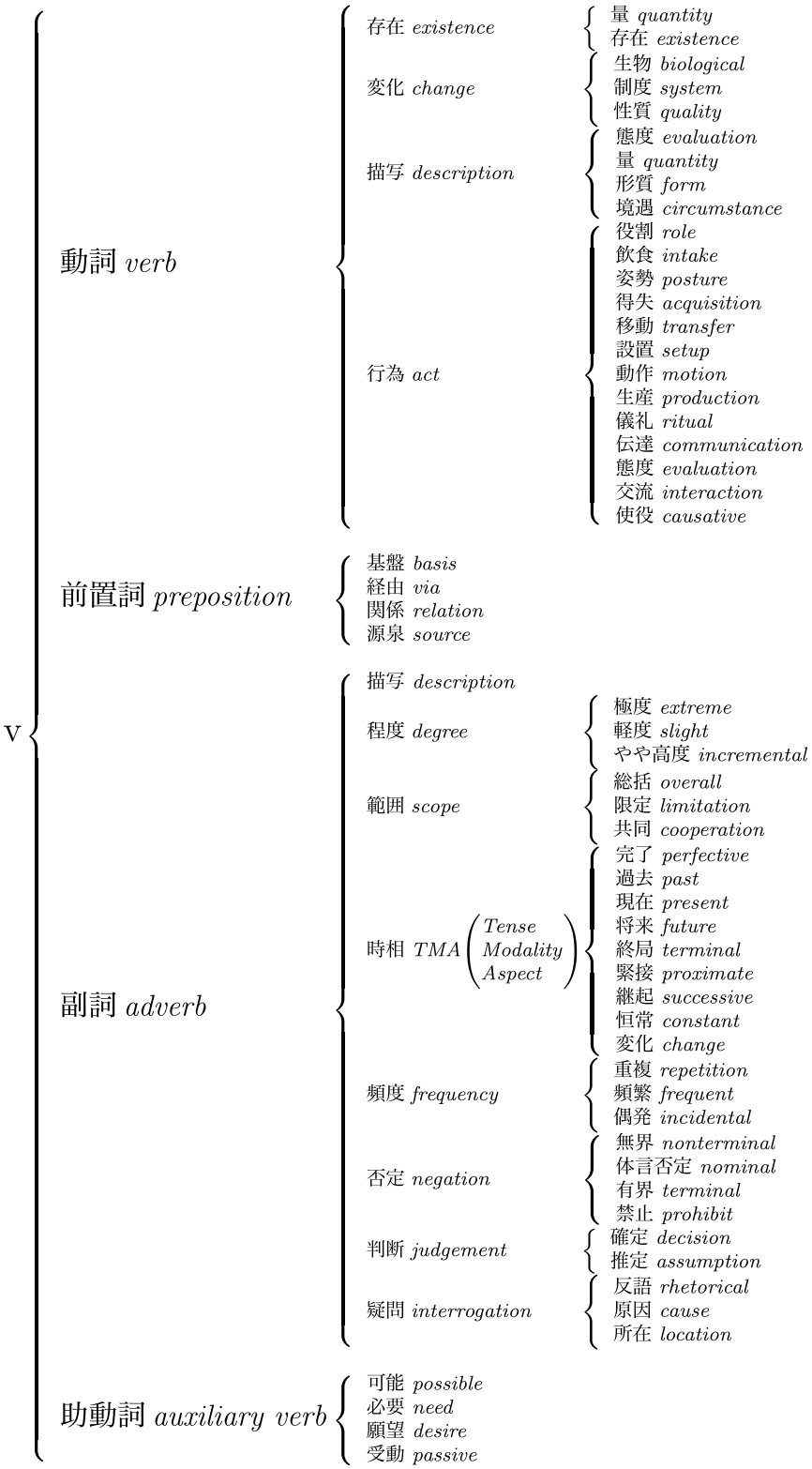


図 3: 形態素解析に特化した古典中国語品詞体系

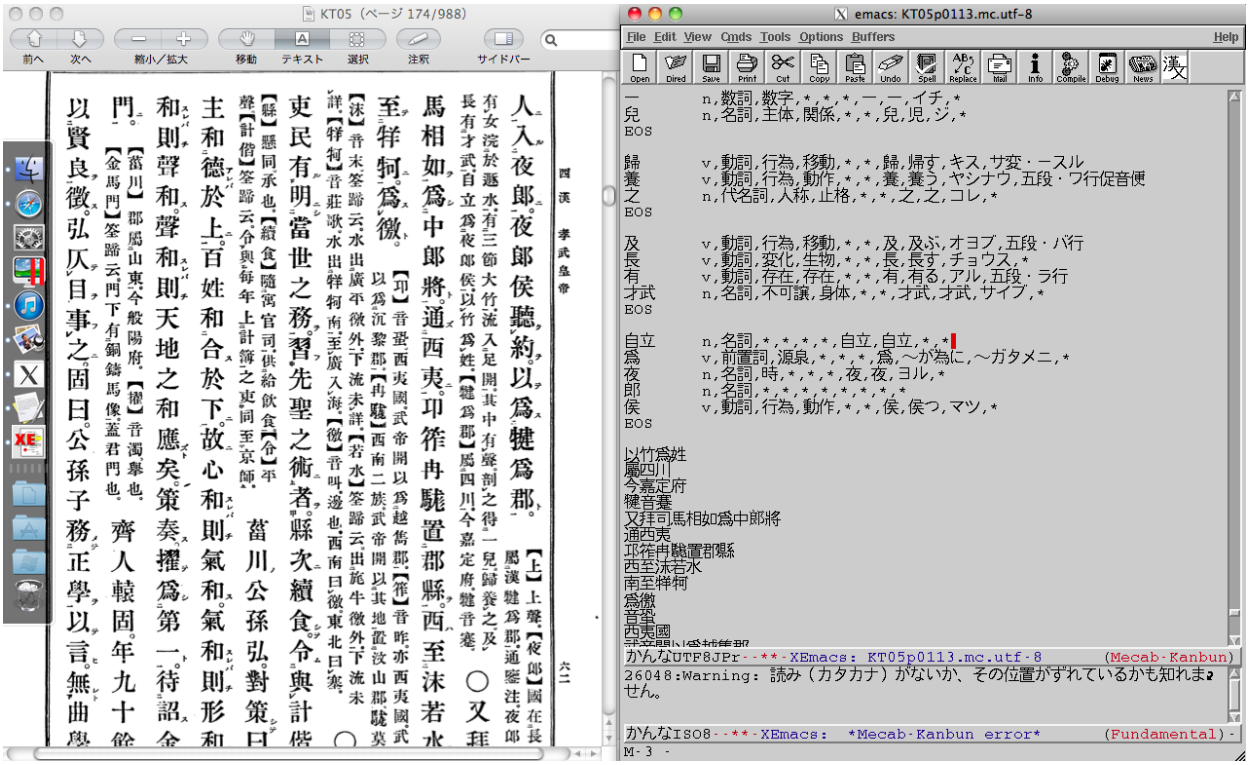


図 4: 漢文コーパス入力ツール

れらに加え、44 種類の意味素性と、88 種類の小素性を定義し、形態素解析の結果として得られる各単語を、意味の面からも捉えやすいよう工夫した。また、この新しい品詞体系による MeCab 漢文辞書を作成すると同時に、MeCab 漢文コーパスにもフィードバックし、全体として新しい品詞体系で、MeCab による漢文の自動形態素解析がおこなえるようにした。

この形態素解析システムで、高校教科書の漢文例や、三国志呉書列伝などの白文を形態素解析してみたところ、まずまずの好結果が得られた [5]。ほぼ全ての白文を単語に切ることが可能となった上に、各単語の品詞もかなり正確に当てることができるようになったのである。

3　漢文コーパスの Linked Data 化

われわれの古典中国語品詞体系と MeCab 漢文コーパスを効果的に管理し、さらには品詞体系のリファクタリングをおこなうべく、われわれは、MeCab 漢文コーパスの Linked Data 化をおこなった [9]。

具体的には、品詞体系の大品詞・品詞・意味素性・小素性の全てを品詞オブジェクトとし、MeCab 漢文コーパスに対しては、見出しオブジェクト (語)、形態素オブジェクト、文オブジェクトの 3 つを準備した。見出しオブジェクトと形態素オブジェクトの間は、対応する品詞オブジェクト (小素性) によってリンクする。形態素オブジェクトは、それを含む文オブジェクトに「用例」としてリンクする。さらに、見出しオブジェクトが 1 文字から構成される場合は、文字オブジェクト (CHISE 文字オントロジー) とリンクする。例として、「自立爲夜郎侯」に関するオブジェクトとリンクを、図 5 に視覚的に示す。ただし、図 5 は、オブジェクトとリンクを概念的に示したものであり、あくまで全体のごく一部であることに注意されたい。

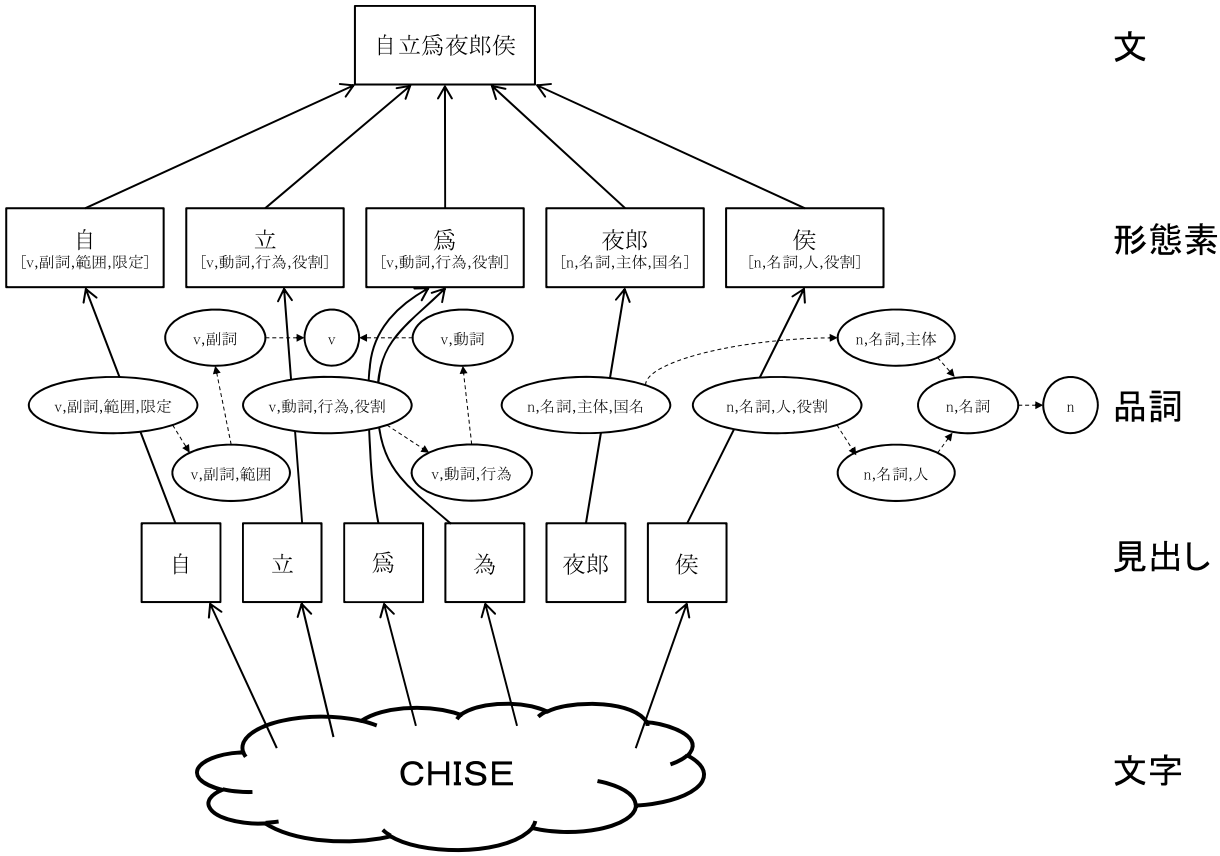


図 5: 「自立爲夜郎侯」の Linked Data 概念図

さらに、これらの Linked Data を WWW 上に実装し、各オブジェクトとリンクを一望できるシステムを実現した [10]。このシステムによって、ある品詞オブジェクトに関する形態素オブジェクトが全て一望できるようになり、品詞体系の効率的なリファクタリングが可能となった。また、ある見出しオブジェクトに関する品詞オブジェクトも一望できるようになった。たとえば「左右」という見出し語に対しては、「n, 名詞, 固定物, 関係」すなわち位置関係を表す場合と、「n, 名詞, 人, 役割」すなわち官職を表す場合があり、それぞれの用例を簡単に辿れるようになった。このシステムを実現したことで、われわれは、漢文の固有表現抽出への手がかりを摑むことができたのである。

4 漢文の固有表現抽出

MeCab 漢文コーパスを用いたさらなる応用として、われわれは、漢文における固有表現の自動抽出に挑戦してみることにした。具体的には、安岡を研究代表者とし、共同研究班の班員全員を研究分担者として、2013 年 4 月から 3 年間、科学研究費補助金基盤研究 (B) 25280122『品詞素性情報つき古典漢文コーパスの発展的応用』の研究助成を受け、漢文における地名・官職・人名の自動抽出に挑戦した。

4.1 地名の自動抽出

漢文での地名を自動抽出する、という目標に向け、われわれは、それまでに作成してきたMeCab 漢文コーパスを洗い直してみた。特に、われわれの新しい品詞体系において「n, 名詞, 固定物, 地名」あるいは「n, 名詞, 主体, 国名」に分類されている形態素オブジェクトと、その形態素オブジェクトを含む文例を見直してみた。この結果、われわれが辿り着いたのが、「2 文字の地名には地名以外の用例はない」という仮説だった。

この仮説に基づき、われわれは「2 文字の地名」の地名以外の用例を、MeCab 漢文コーパスに対して、サンプリング調査してみた。そうしたところ、そのような地名以外の用例は、どの「2 文字の地名」においても 10%未満だった。しかも、それら 10%足らずの用例も「n, 名詞, 固定物, 地形」など、山や川の名前を例文入力グループが地形だとみなしたものが大多数で、これらを仮に地名だとみなしても大した問題は起こらない。「2 文字の地名には地名以外の用例はない」という仮説は、少なくとも 90%の確率で当たっており、地名の自動抽出という観点からは、採用するに値する。

この結論に基づき、われわれは、MeCab 漢文コーパスから抽出した「2 文字の地名」を、そのまま MeCab 漢文辞書に追加した。また、3 文字以上の地名は、その多くが「○○府」や「○○縣」の形を取るものだったが、同様に MeCab 漢文辞書に追加した。

では、「1 文字の地名」は、どうなのか。たとえば「渭」のように、地名用例しかないような「1 文字の地名」に関しては、そのまま MeCab 漢文辞書に追加すればよい。しかし、たとえば「夏」という形態素は、王朝名としての「夏」かもしれないし、季節としての「夏」かもしれない。あるいは「莫」という形態素は、地名用例はむしろ少数で、大多数の用例が「v, 副詞, 否定, 禁止」である。もし、「莫」を無理矢理に地名だとみなすような処理をおこなうと、「v, 副詞, 否定, 禁止」であるべき「莫」を、誤って「n, 名詞, 固定物, 地名」だと処理してしまう危険性がある。その場合、後続の動詞にも悪影響が及ぶので、文法上のミスとしては致命的である。そのようなミスは、絶対に避けなければならない。

この問題に対し、われわれは、たとえ「1 文字の地名」を全て MeCab 漢文辞書に追加したとしても、MeCab 漢文コーパスを十分に準備すれば、そのようなミスは形態素解析において発生しないだろう、という希望的観測を持ってみることにした。「2 文字の地名」という巨大な用例による接続確率 (裏を返せば非接続確率) が効いてくるはずで、それによって「1 文字の地名」も正しく認識されるはずだ、という甘い予想を立てたわけである。

もちろん、この予想がうまくいくためには、他の地名用例コーパスも含め、できるだけ多くの地名用例コーパスが必要な上に、対抗用例コーパスも十全に収録しておかねばならない。たとえば「莫」であれば、「n, 名詞, 固定物, 地名」の「莫」も、「v, 副詞, 否定, 禁止」の「莫」も、いずれも MeCab 漢文辞書に含まれている必要があるし、「莫」の副詞用例コーパスも十全に収録しておかねばならない。また、地名用例コーパスや対抗用例コーパスに加え、それら以外のコーパスも、バランスよく収録しておく必要がある。この目標のために、われわれは、それまでに入力していた約 46000 文の MeCab 漢文コーパスから、複数の入力者による分析結果が品詞レベルで完全に一致した用例 (約 2000 文、地名を約 400 語収録) を、本手法の学習用コーパスとして用いることにした。

この手法により、われわれの形態素解析システムは、たとえば「莫滅莫」という (かなり人工的な) 漢文を

莫　　v, 副詞, 否定, 禁止,*,*, 莫, 莫し, ナシ,*
滅　　v, 動詞, 変化, 制度,*,*, 滅, 滅す, ホロボス, 五段・サ行
莫　　n, 名詞, 固定物, 地名,*,*, 莫, 莫, バク,*

「莫を滅すなかれ」と正しく処理できるようになった。また、この手法を定量的に評価すべく、地名を加えない MeCab 漢文辞書との比較をおこなった [12]。この結果として、できる限り多くの地名を MeCab 漢文辞書に追加する手法は、地名を含む漢文の認識精度を高めると同時に、地名を含まない漢文には悪影響がない、という形で、本手法の有効性が確認された。

4.2　官職の自動抽出

漢文における官職を自動抽出する際も、文字数の短い官職であれば、地名と同様の手法が効果的だった。実際、MeCab 漢文辞書と MeCab 漢文コーパスを十全に準備することで、たとえば「上下左右」の「左右」と、「引置左右」の「左右」を

上下　n, 名詞, 固定物, 関係,*,*, 上下, 上下, ジョウゲ,*
左右　n, 名詞, 固定物, 関係,*,*, 左右, 左右, サユウ,*

引　　v, 動詞, 行為, 動作,*,*, 引, 引く, ヒク, 五段・カ行イ音便
置　　v, 動詞, 行為, 設置,*,*, 置, 置く, オク, 五段・カ行イ音便
左右　n, 名詞, 人, 役割,*,*, 左右, 左右, サユウ,*

という形で正しく見分けることは、われわれの形態素解析システムでは既に可能となっている*。

その一方、複数の形態素から構成される (ように見える) 官職もあり、これがわれわれを悩ませた。以下に、いくつかの典型例を示す。

- 丞
  「○○丞」の形を取る名詞は、ほぼ全て官職とみなせる。しかしながら、その形態素解析処理は問題を孕んでいる。たとえば「御史中丞」を一つの形態素だとみなしてしまうと、「右御史中丞」や「知御史中丞」をうまく処理できない。「右御史臺中丞」となると、もうどうしていいかわからない。また、「右」は必ずしも最初に付加されるとは限らず、「尚書右丞」「尚書左丞」のような例もある。これらに加え、「湖州長城丞」や「長沙縣丞」のように地名との複合が起こる場合もあって、混沌を極める。

- 郎中
  「○○郎中」の形を取る名詞は、まず間違いなく官職である。これらのうち、「兵部郎中」や「司勳郎中」のように、部署名や他の官職との単純な複合は、まだ何とか処理できる。しかしこれが、「兵部左司郎中」や「尚書司勳郎中」という形で複合すると、もはや形態素解析の手に負えない。

---

*「引」と「置」は、本来は「v, 動詞, 行為, 役割」であるべきだ。しかし、現状のわれわれのシステムは、この例文において、「引」を「v, 動詞, 行為, 動作」、「置」を「v, 動詞, 行為, 設置」だと読んでしまう。動詞類も、さらに鍛える必要がある、ということだろう。

- 判〜事、知〜事、〜従事
  「判○○事」「知○○事」「○○従事」の形を取る名詞は、かなりの確率で官職である。しかしながら、形態素解析の立場からすると、「判」「知」「従」はいずれも動詞とみなすべき形態素であり、これが問題を複雑にしている。たとえば「知民事」は、通常は「民事を知る」という文であって、官職ではない。一方「知政事」は官職である。あるいは「知吏部尚書事」は官職だが、内部に他の官職である「吏部尚書」を含んでしまっている。

複数の形態素から構成される官職は、形態素処理の後に「組み上げ処理」をおこなうことで、抽出可能だと考えられる。しかしながら、この「組み上げ処理」は、たとえば「丞」と「郎中」と「事」とで、全く異なる処理をおこなわざるを得ない。つまり、官職中に使われている文字ごとに異なった処理が必要で、それぞれをバラバラに力業で組み上げるしかない、ということである。

4.3 人名の自動抽出

漢文における人名の自動抽出に向けて、われわれは、MeCab 漢文コーパスを洗い直し、「n, 名詞, 人, 姓氏」「n, 名詞, 人, 名」に分類されている形態素オブジェクトと、その形態素オブジェクトを含む文例を見直してみた。その結果、「n, 名詞, 人, 姓氏」については、地名抽出と同様の手法が有効だ、との感触が得られた。しかし、「n, 名詞, 人, 名」については、他の用例とのバッティングが、奇妙な方法で回避されていることが判明した。具体例として、『十八史略』巻之二に現れる「李斯」という人名に関して、われわれが得た知見を、以下に述べる。

『十八史略』巻之二には、「斯」という文字が、全部で 16 例、出現する。これらのうち 6 例は、「李斯」という形で出現することから

```
李　　n, 名詞, 人, 姓氏,*,*, 李, 李, リ,*
斯　　n, 名詞, 人, 名,*,*, 斯, 斯, シ,*
```

であることは確実であり、実際の形態素解析においても、そう処理できる。問題は残る 10 例である。これら 10 例は「斯」が単独で出現するのだが、われわれの判断では、最初の 9 例は全て「n, 名詞, 人, 名」であり、最後の 1 例だけが「n, 代名詞, 指示,*」なのである。具体的には、「李斯」の話が続いている間は、ずっと「斯」は特定の人名である「李斯」を指しており、その後「李斯」が出てこなくなって、かなり文章が進んでから、やっと代名詞の「斯」がたった 1 例だけ出現する。つまり、「李斯」の話が続いている間は、話がややこしくならないよう、代名詞の「斯」の使用をあえて避けているわけである。

このような形で、「ストーリー」全体における用字の分布に異常が見られる場合、それが人名を指している可能性が高い、ということは推定できた。「斯」の例で言えば、曖昧語となりうる「斯」の曖昧性を下げるために、代名詞としての「斯」を使わない、という形で『十八史略』巻之二における用字分布が変わってしまっている。しかしながら、この推定を、人名の自動抽出にまで結びつけるような手法は、われわれには開発し得なかった。というのも、「ここで斯が出てきたなら、それは李斯であって、代名詞じゃないよな」ということを理解するには、本質的には「ストーリー」の理解が必要だからである。現状のわれわれの力不足を、痛感する限りである。

5　おわりに

ほぼ 8 年間にわたる、われわれの苦闘の歴史を、本稿では概観した。この苦闘の結果、漢文の形態素解析という局面において、特に名詞まわりの処理については、かなり良い成果が得られたと信じる。

ただし、それはあくまで定性的な側面であり、かならずしも定量的な評価が得られたわけではない。もちろん、F 値[†]による評価は数次に渡っておこなった [2, 5, 12] のだが、どうも納得がいかないのだ。われわれとしては、われわれが構築している形態素解析システムにおいて、そのインテリジェンス (というか、かしこさ) を評価したいのだが、F 値は単純に間違いをチェックするだけである。具体的には、漢文コーパスを増やしていっても F 値はあまり変化しないのだが、その結果をわれわれが読む限りでは、F 値は変わらなくても、やはり少しずつ「かしこく」なっているのだ。漢文の文法上、スジのいい間違いとスジの悪い間違いが現実には存在するのだが、それらの差異をうまく引き受けてくれるような評価尺度を、われわれは未だ見つけ得ていないのである。

その意味では、われわれの闘いは、まだまだ続くということなのだろう。2016 年 4 月より、新たに京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの実証研究」を発足する予定である。われわれの今後の研究の発展に、ぜひ期待されたい。

[†]適合率を $\frac{R}{N}$、再現率を $\frac{R}{C}$ (R:処理結果中の正しい適合数、N:処理結果数、C:対象全体における適合すべき数) とおく時、F 値は適合率と再現率の調和平均すなわち $\frac{2\times適合率\times再現率}{適合率+再現率}$。

関連発表

[1] 守岡知彦: MeCab を用いた古典中国語の形態素解析の試み, 情報処理学会研究報告, Vol.2008-CH-79 (2008 年 7 月), pp.17-22.

[2] 守岡知彦: MeCab を用いた古典中国語形態素解析器の改良, 情報処理学会研究報告, Vol.2009-CH-84 (2009 年 10 月), No.3, pp.1-5.

[3] Tomohiko Morioka: A Prototype of a Classical Chinese Morphological Analyzer based on MeCab, Osaka Symposium on Digital Humanities 2011 (September 2011), p.36.

[4] 守岡知彦: 古典中国語形態素コーパス編集システムの開発, 東洋学へのコンピュータ利用, 第 23 回研究セミナー (2012 年 3 月), pp.75-83.

[5] 山崎直樹, 守岡知彦, 安岡孝一: 古典中国語形態素解析のための品詞体系再構築, 人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集 (2012 年 11 月), pp.39-46.

[6] 「東アジア古典文献コーパスの研究」文献ログ (2008 年 4 月〜2013 年 3 月)
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2013.html

[7] Tomohiko Morioka, Christian Wittern, Koichi Yasuoka, Naoki Yamazaki: A Study of Linguistic Analysis for Classical Chinese Texts, Proceedings 2013 International Conference on Culture and Computing (September 2013), pp.143-144.

[8] 「東アジア古典文献コーパスの研究」共同研究班報告, 東方學報, 第 88 冊 (2013 年 12 月), pp.292-287.

[9] 守岡知彦: 古典中国語形態素コーパスの Linked Data 化の試み, 人文科学とコンピュータシンポジウム「じんもんこん 2013」論文集 (2013 年 12 月), pp.187-194.

[10] 守岡知彦: 比較的最近の CHISE, 東洋学へのコンピュータ利用, 第 25 回研究セミナー (2014 年 3 月), pp.33-46.

[11] Koichi Yasuoka, Naoki Yamazaki, Christian Wittern, Yoshihiro Nikaido, Tomohiko Morioka: A Morphological Analysis of Classical Chinese Texts, Digital Humanities 2014 (July 2014), pp.410-412.

[12] 安岡孝一, 守岡知彦, Christian Wittern, 山崎直樹, 二階堂善弘, 鈴木慎吾: 古典中国語形態素解析による地名の自動抽出, 人文科学とコンピュータシンポジウム「じんもんこん 2014」論文集 (2014 年 12 月), pp.63-68.

[13] 「東アジア古典文献コーパスの応用研究」ログ (2013 年 4 月〜2016 年 3 月)
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2016.html